Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 32 (2018) 416–420

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Steering Committee for Transport Survey Conferences

Workshop Synthesis: Data analytics and fusion in a world of multiple sensing and information capture mechanisms

Elisabetta Cherchi[a]*, Chandra Bhat[b]

[a]Newcastle University - School of Engineering, Cassie Building, Newcastle upon Tyne, NE1 7RU, UK
[b]Department of Civil, Architectural and Environmental Engineering - The University of Texas at Austin, Austin, Texas 78712

Abstract

This paper is a synthesis of the discussions and ideas that were generated during the workshop on "Data Analytics and Fusion in a World of Multiple Sensing and Information Capture Mechanisms" at the 2017 Travel Survey Methods Conference in Esterel (Canada). The workshop explored and discussed a new landscape in which data from multiple sensing and information capture mechanisms may be gainfully combined to obtain richer, more comprehensive, and more representative information regarding human mobility patterns as well as human perceptions, attitudes, and lifestyle choices. The workshop discussed the exciting possibilities, some investigative and predictive analytics pathways forward in terms of methods, and the research challenges in this emerging landscape of data science.

Keywords: travel behavior data, heterogenous data sources, high-dimensional data, causality and association in relationships.

1. Introdution

This workshop focused on the emerging data landscape in which data from multiple sensing and information capture mechanisms may be gainfully combined to obtain richer, more comprehensive, and more representative information regarding human mobility patterns as well as human perceptions, attitudes, and lifestyle choices. In particular, there is an opportunity to combine data from multiple capture mechanisms in ways that compensate for the limitations of any single source capture mechanism, and harness the collective strengths across multiple capture

* Corresponding author. +44 (0) 191 208 3501
E-mail address: elisabetta.cherchi@ncl.ac.uk

10.1016/j.trpro.2018.10.059

mechanisms. Such data analytic and fusion strategies require the ability not only to deal with data from multiple sources, but also potentially highly noisy, heterogeneous, and high-dimensional data with complex interdependencies.

The workshop started with four presentations (Ellison and Lovelace, 2017; Hilgert et al., 2017; Bourbonnais and Morency, 2017; Katranji et al., 2017) that set the stage for the core discussions regarding the challenges as well as the opportunities as we move into a new era of multiple-source data collection and data analysis. The four presentations related to (a) augmenting traditional surveys with ubiquitous "big" data from non-survey capture mechanisms, (b) combining objective performance data with opinion surveys, (c) fusing data from commercial vehicle GPS and roadside intercept data and (d) validating travel behaviour from smartcard data.

Following the discussion at the workshop, this paper, in Section 2, discusses the different sources of data available in the emerging new data landscape. Section 3 focuses on the new challenges of this new landscape, as well as reports on a discussion on how to store the new data to facilitate easy fusion. Section 4 summarises the main conclusions and the research priorities that emerged from the workshop.

2. Different sources of data

The concept of data fusion is not new. The most notable example is probably the use of revealed (RP) and stated preference (SP) data that have been used since the early 90s (Ben-Akiva and Morikawa, 1990; Bradley and Daly, 1997; Louviere et al., 2000; Brownstone et al., 2000; Bhat and Castelar, 2002; Cherchi and Ortúzar, 2002; Bhat and Castelar, 2006) to overcome the limitations of RP data (lack of attribute variability, correlation between attributes, lack of data on alternatives not in the market) and the limitations of the SP data that do not represent the real market. More recently, attitudinal data or opinion data have also been used jointly with the more traditional travel data to capture behavioural processes and more parsimoniously accommodate heterogeneities in preferences and taste sensitivities across individuals (see, for example, Ben-Akiva et al., 2002, Daziano and Bolduc, 2013, Bhat and Dubey, 2014, Bhat, 2015).

Despite the long tradition of combining data from different sources, the recent and upcoming availability of new emerging data sources open new opportunities and new challenges (e.g. Bayart et al., 2009) These emerging data include those collected by geographic positioning units (GPS) (Stopher et al., 2015), smartcards, social media sources (twitter, trip advisor, Facebook), smartphone (Ellison et al., 2014), vehicle data (Kuhnimhof et al., 2011), travel transactional data (e.g. car sharing), consumption transactional data, sensor data (from Bluetooth, videos, Mac address-based sources, drones, and remote sensing/satellite images). Differently from the traditional data, emerging data are characterised by voluminous amounts of near-real time data, and they do not require sampling, being able to reach almost the entire population (e.g. when using telephone and bank services) or at least the entire population within groups (e.g. when using smart cards, toll roads etc.). At the same time, these emerging data are also typically coarser and more sectorial than traditional data. Table 1 provides a synthesis of the main differences between traditional and emerging data.

The workshop addressed the question of the benefits of acknowledging and accommodating the new emerging data landscape. In doing so, participants felt the need to identify the unique (as well as the ubiquitous) new landscape of data capture in terms of obtaining information on consumer characteristics, habits, attitudes, preferences, and activity-travel patterns. It was generally agreed that the difference in the use of traditional and emerging data is in the nature of the prediction process under consideration. In particular, traditional data are grounded into basic behavioural concepts and typically have been collected to be used in policy and proactive analysis (causality is the main interest). On the other hand, the new emerging data usually (but not always) do not include a data collection design based on behavioural considerations and are typically used to reveal associations among data rather than causality. At the moment, much of our research and applied work has focused on causality rather than association, and therefore on traditional data. But this is probably because we are only recently discovering new ways to extract causality from new emerging data, and much more understanding and method development is needed in this area. Some works are trying to infer individual behaviour from big data. Eventually emerging data can and may substitute completely for traditional data, but, for the moment, it is only reasonable that we continue to collect and use data from both traditional and emerging sources.

Table 1. Differences between traditional and emerging data.

Traditional data	Emerging data
strategic sample	almost the entire population
respondent stated (RP/SP)	"true" observed data †
fine / demographics (disaggregate)	coarse (aggregate)
multidimensional (many small pieces of information for the same person) on mobility and individual	unidimensional (more sectorial) focused on mobility
small sample/easy to process	voluminous/more difficult to process
clarity in sharing / privacy / access	fuzziness in sharing / privacy / access (not all data are clearly associated to a specific person)

3. Fusion, challenges, and data warehousing

The differences in mobility patterns as implied by different data capture mechanisms, and the potential to fuse data from multiple data sources to extract useful information for travel analysis, pose new challenges that, during the workshop, where identified as:

- Privacy issues with being tracked all the time. Not all data are clearly associated to a specific person. Maybe over time, barriers can be relaxed;
- How to merge the data and validate fusion methods;
- Methods: some work exists, but more papers and more research efforts are needed.

Another key problem that emerged during the workshop related to how to "house" data from diverse sources in ways that make it easy for data processing and compilation. We need to have a unique place where all the data collected are stored and available to researchers. Often new emerging data are collected by different agencies, and the challenge is then to develop a standardization mechanism that allows fusion. Issues for further analysis and protocol development include which type of data to store and make available, who should do what, and who should have access to what. Currently, the diversity of protocols and formats in which data is collected (e.g. Europe is very heterogeneous, USA depends on States) makes standardization across the board a real challenge. We probably need to move in incremental steps, standardizing only some type of information, while leaving others flexible. At the same time, it is important to think about protocols and procedures for broader standardization.

While the transport community has only relatively recently started to use new emerging data sources, other communities (such as neuroscience, physician and machine learning communities) have more experience from which we may be able to learn, while also continuing to maintain our emphasis on behavioural concept-oriented data collection designs and processing.

4. Conclusion

The objective of this workshop on "Data Fusion: Needs and Challenges in a New Transportation Data Landscape" was to explore the potentiality of the emerging data coming from multiple different sources and how this can be efficiently combined to obtain richer, more comprehensive, and more representative information regarding human mobility patterns, perceptions, attitudes, and lifestyle choices. The workshop generated a lively discussion and many fruitful ideas. Differences in how we should proceed emerged among the participants of the workshop, due to the

† During the workshop, the definition of "true" data to highlight the fact that emerging data are not "filtered" by the respondents when interviewed. "True" was not used with the meaning of "free of bias" as all systems have their own limitations which might be technical, or data transformation, or related to user misuse (fraud or multiple validation for example for smart card data). It also happens that providers of these data do not always give access to the original data, but sometimes only to the processed data.

diversity in background and the interesting challenges that lie ahead. A convergence, however, was reached in terms of research needs, identified as follows:

- How to optimise resources/investments when we combine traditional/new data sources?
- How do we store data in an efficient/standard way and make them widely available?
- How do we use these merged sources of data for informing policies? What are the sample size requirements of different data sources and what procedures may be effective to merge data?
- What information can we merge (companies do it)?
- Perhaps a good way to merge data would be to undertake a meta-analysis on different types of fusion and their advantages and limitations. Such a systematic assessment of the usefulness of alternative fusion methods is imperative. Also, more efforts are needed on what new data is easy to access and combine?
- How to bridge the gap between data driven and theory driven approaches to travel behaviour analysis?
- How can we be proactive despite the challenges, and how can we interact more with researchers in other fields (machine learning, robotics, artificial intelligence)?

Acknowledgements

The authors are grateful to all the workshop participants for their valuable contributions in discussing the new emerging data landscape and identifying future directions to make the most of the uniqueness and heterogeneity of the different data sources. In particular we thank: Farzad Alemi (USA), Pierre-Leo Bourbonnais (Canada), Ricardo Daziano (USA), Elodie Deschaintres (Canada), Richard Ellison (Australia), Tim Hilgert (Germany), Ryoji Ishii (Japan), Martin Kagerbauer (Germany), Tobias Kuhnimhof (Germany), Nadav Levy (Israel), Tianyang Lin (Canada), Patrick Loa (Canada), Henry Mlotsa (South Africa), Marcela Munizaga (Chile), Fouad Hadj Selem (France), Gabriel Sicotte (Canada), Pedro Szasz (Barsil).

References

Bayart, C., Bonnel, P., Morency, C., 2009. Survey Mode Integration and Data Fusion: Methods and challenges. In Transport Survey Methods, Keeping up with a Changing World. Eds. P. Bonnel, M. Lee-Gosselin, J. Zmud, J.L. Madre. Emerald. Chapter 34, 587-611.

Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bunch, D.S., 2002. Hybrid choice models: Progress and challenges. Marketing Letters, 13(3), 163-175.

Ben-Akiva, M.E., Morikawa, T., 1990. Estimation of travel demand models from multiple data sources. 11th International Symposium on Transportation and Traffic Theory.Yokohama, Japan.

Bhat, C., Castelar. S., 2002. A unified mixed logit framework for modelling revealed and stated preferences: Formulation and application to congestion pricing analysis in the San Francisco Bay Area. Transportation Res. 36B 593–616.

Bhat, C., Sardesai. R., 2006. The impact of stop-making and travel time reliability on commute mode choice. Transportation Res. 40B 709–730.

Bhat, C.R., Dubey, S.K., 2014. A new estimation approach to integrate latent psychological constructs in choice modeling. Transportation Research Part B, 67, 68-85.

Bhat, C.R., 2015. A New Generalized Heterogeneous Data Model (GHDM) to Jointly Model Mixed Types of Dependent Variables, Transportation Research Part B, 79, 50-77.

Bradley, M.A., Daly A.J., 1997. Estimation of logit models using mixed stated preference and revealed preference information. In: Stopher PR & Lee-Gosselin M (eds) Understanding Travel Behaviour in an Era of Change. Oxford: Pergamon.

Bourbonnais, P.L., Morency, C., 2017. A Robust Datawarehouse as a Requirement to the Increasing Quantity and Complexity of Travel Survey Data. 11th International Conference on Transport Survey Methods, Esterel, Canada.

Brownstone, D., Bunch, D., Train. K., 2000. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. Transportation Res. 34B 315–338.

Chigoy, B., Hard, E., Martin, M., Green, L., 2017. Passive Data: The Other 50% of the Work. 11th International Conference on Transport Survey Methods, Esterel, Canada.

Cherchi, E., Ortúzar, J. de D., 2002. Mixed RP/SP models incorporating interaction effects: Modelling new suburban train services in Cagliari. Transportation 29 371–395.

Daziano, R.A., Bolduc, D., 2013. Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model. Transportmetrica A: Transport Science, 9(1), 74-106.

Ellison, R., Ellison, A., Greaves, S., Standen, C., 2014. Harnessing smartphone sensors for tracking location to support travel data collection. 10th International Conference on Transport Survey Methods, Leura, Australia.

Ellison, R., Lovelace, R., 2017. Augmenting travel surveys with Big Data. 11th International Conference on Transport Survey Methods, Esterel, Canada.

Hilgert, T., Von Beren, S., Kostorz, N., Kagerbauer, M., Vortisch, P., 2017. Does travel behavior of people using mobility apps differ? Findings from a market analysis in Germany. 11th International Conference on Transport Survey Methods, Esterel, Canada.

Katranji, M., Sanmarty, G., Kraiem, S., Hadj Selem, F., 2017. Inferring human mobility patterns from census data. 11th International Conference on Transport Survey Methods, Esterel, Canada.

Kuhnimhof, T., Ottmann, P., Zumkeller, D., 2011. Adding Value to Your Data: Analysis of Travel Expenses Based on Trip Diary and Enriched Odometer Reading Data 9th International Conference on Transport Survey Methods, Termas de Puyehue, Chile.

Kuhnimhof, T., Weiss, K., 2017. Vehicle cost imputation in travel surveys: Gaining insight into the fundamentals of (auto-) mobility choices. 11th International Conference on Transport Survey Methods, Esterel, Canada.

Louviere, J.J., Hensher D.A., Swait, J.D., 2000. Stated Choice Methods: Analysis and Application. Cambridge: Cambridge University Press.

Stopher, P., Shen, L. Liu, W., Ahmed, A., 2015. The Challenge of Obtaining Ground Truth for GPS Processing. Transportation Research Procedia, 11, 206-217.